File Harvest

Targeted, Legal Crawling and Downloading of Online Media

Chad Sowald, Computer Science and Engineering, Ohio State University
Paul Sivilotti, Advisor

Goal

Download media found while crawling web pages to eliminate the need for users to manually browse the same pages and manually download the same media.

Results

File Harvest, a Windows application, finds and downloads media – videos, music, images, etc… – served statically through HTML and dynamically through the HTTP protocol.

Importance

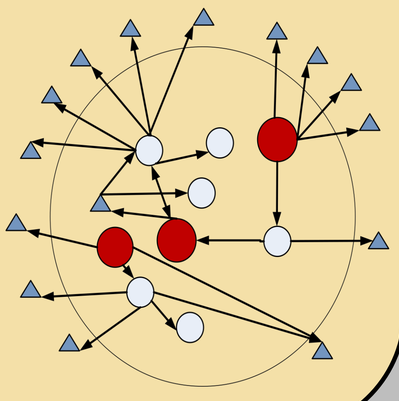
Internet users waste time manually browsing and downloading media from websites. Users need a way to automatically download this media and view it offline.

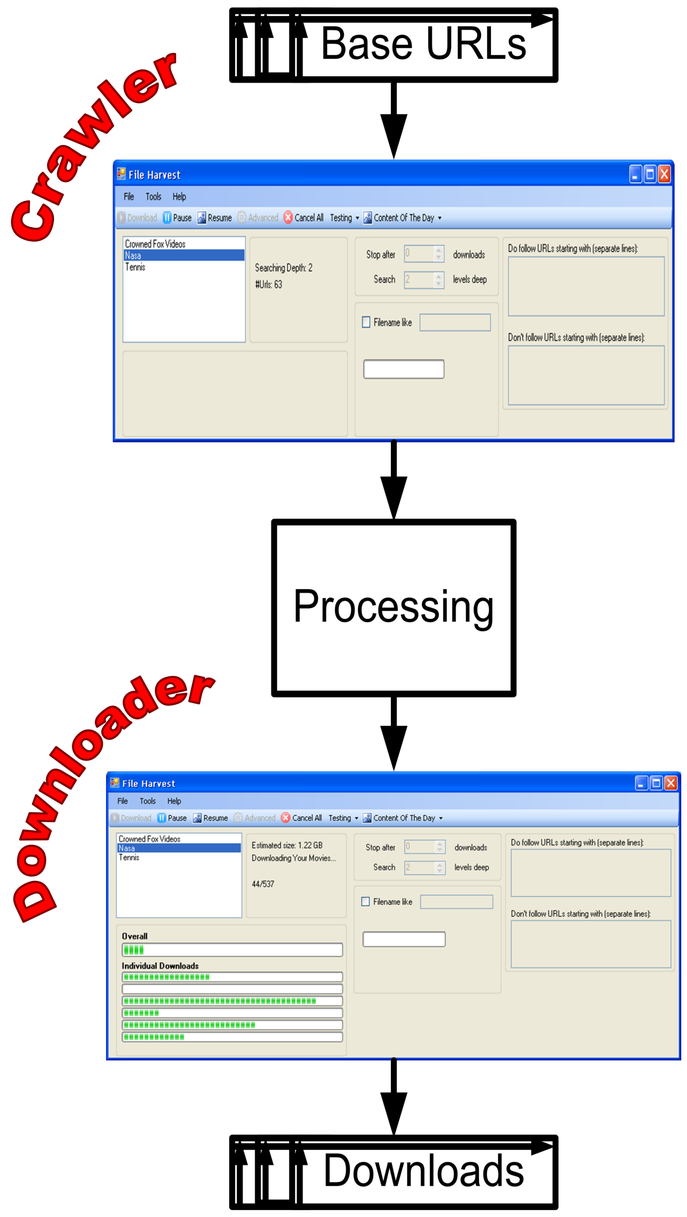
Going Forward

Support for other popular media delivery protocols such as the Real Time Messaging Protocol (RTMP) to capture media from a wider range of web sites.

Web Crawling

File Harvest works by automatically following hyperlinks on web pages to arrive at other pages. At each page, the crawler searches for media the user has configured File Harvest to download. In the graph to the right, large red circles indicate the web pages File Harvest starts crawling from, the smaller circles are web pages found during crawling and the triangles are files that are downloaded. All nodes within the very large circle are crawled—including any triangles. This is because File Harvest can download a web page and also crawl it.





Users configure File Harvest before starting a crawling and downloading session. The session produces downloads that the user can view offline.

MIME Types

MIME, or Multipurpose Internet Mail Extensions, are a way for web servers to communicate the type of content located at a particular URL. File Harvest uses MIME types to decide what files to download and which files it can crawl.

MIME Class	Description
application	Typically, binary files. Examples include EXE and DOC files.
audio	Audio and sound files. Examples include MP3 and WAV files.
image	Image and picture files. Examples include JPEG and BMP files.
text	Text files. Human readable. Examples include HTML and TXT files.
video	Video files. Examples include AVI, WMV, QT, and FLV files.

File Harvest can capture dynamically served media over the HTTP protocol because it proxies the Windows HTTP handler, WinINET.